Why are these similar?
Investigating item similarity types in a large Digital Library

Aitor Gonzalez-Agirre
University of the Basque Country
Informatika Fakultatea, Donostia 20018, Basque Country
Phone: +34 943 015 019
Fax: +34 943 015 810
aitor.gonzalez-agirre@ehu.es

Nikolaos Aletras
University of Sheffield
Regent Court, 211 Portobello, Sheffield, UK
Phone: +44 114 222 1921
Fax: +44 114 222 1810
n.aletras@sheffield.ac.uk

German Rigau
University of the Basque Country
Informatika Fakultatea, Donostia 20018, Basque Country
Phone: +34 943 015 019
Fax: +34 943 015 810
german.rigau@ehu.es

Mark Stevenson
University of Sheffield
Regent Court, 211 Portobello, Sheffield, UK
Phone: +44 114 222 1921
Fax: +44 114 222 1810
mark.stevenson@sheffield.ac.uk

Eneko Agirre (corresponding author)
University of the Basque Country
Informatika Fakultatea, Donostia 20018, Basque Country
Phone: +34 943 015 019
Fax: +34 943 015 810
e.agirre@ehu.es

Abstract

We introduce a new problem, identifying the type of relation that holds between a pair of similar items in a digital library. Being able to provide a reason why items are similar has applications to recommendation, personalisation and search.

We investigate the problem within the context of Europeana, a large digital library containing items related to cultural heritage. A range of types of similarity in this collection were identified. A set of 1500 pairs of items from the collection were annotated using crowdsourcing. A high inter-tagger agreement (average 71.5 Pearson correlation) was obtained and demonstrates that the task is well defined. We also present several approaches to automatically identifying the type of similarity. The best system applies linear regression and achieves a mean Pearson correlation of 71.3, close to human performance.

The problem formulation and data set described here were used in a public evaluation exercise, the *SEM shared task on Semantic Textual Similarity. The task attracted the participation of 6 teams, who submitted 14 system runs. All annotations, evaluation scripts and system runs are freely available[1].

1 Introduction

Search engines and digital libraries often allow users to search for similar items, an important function which supports exploratory search (Marchionini, 2006) and sense-making (Hearst, 2009). Users are often provided with similar items in the form of a link from an individual item to a set of others in the collection. For example, Google Scholar[2] and PubMed[3], both digital libraries containing academic publications, provide users with such links. Google Scholar has a link to "Related Articles" and PubMed to "Related Citations". This feature is so important that it is implemented in many open-source search engines, e.g. Lucene and Terrier (Ounis et al., 2006; McCandless et al., 2010).

Similar items are normally identified using word-overlap measures. Following this approach, the similarity of a pair of documents is determined by counting the number of words they have in common, possibly with adjustment for factors such as document length and word frequency (Baeza-Yates and Ribeiro-Neto, 1999; Manning and Schütze, 1999; Jurafsky and Martin, 2009). This approach has the advantage of being robust, straightforward to compute and is useful for identifying pairs of documents describing closely related topics.

However, items in collections can be similar in different ways. For example, two documents in a collection could be considered to be similar if they discuss the same topic or are written in the same style. The ways in which items can be considered similar also varies between collections. In collections of academic publications, such as Google Scholar or PubMed, pairs of citations could be considered to be similar for several reasons including being written by the same authors, citing the same publications, describing the same type of scientific investigation (e.g. a clinical trial or a meta study) or having the same conclusions. In different collections other features may be more relevant for determining whether items are similar.

Existing methods for identifying similar items within collections do not acknowledge that there are different ways in which items can be similar. This paper explores the problem of identifying different types of similarity[4] in a large digital library containing a collection of information about cultural heritage artefacts, Europeana (see Section 2). The nature of the cultural heritage domain makes it appropriate for exploring the typed similarity problem. There are several ways in which the items in cultural heritage can be considered to be similar and identifying them has useful applications, including making recommendations (Resnick and Varian, 1997; Grieser et al., 2007; Bohnert et al., 2009), supporting exploratory search (Marchionini, 2006), personalisation (Bowen and Filippini-Fantoni, 2004; O'Donnell et al., 2001) and (automatic) tour generation (Finkelstein et al., 2002; Roes et al., 2009; Agirre et al., 2013a).

[1] http://ixa2.si.ehu.es/sts

[2] http://scholar.google.com/

[3] http://www.ncbi.nlm.nih.gov/pubmed

[4] We use the term similarity in this paper since it is more commonly used in the research literature. We acknowledge the distinction between similarity and relatedness, and ask raters to judge similarity between items (see later sections). However, the term similarity is used to capture both concepts for simplicity.

In this paper we present the first dataset for the typed similarity problem. The data set contains pairs of Cultural Heritage items from Europeana to which we assigned a scores for a range of similarity types: similar author, similar people involved in the items, similar time period, similar location, similar event, similar subject and similar description. The dataset contains 1500 pairs of items that were manually annotated with those types using crowdsourcing. The annotators assigned a number between 0 (completely unrelated) to 5 (identical) for each type of similarity. The annotations are reliable, as demonstrated by high inter-tagger correlation agreement. In addition, we also developed a system that accurately produces typed similarity judgements, using similarity-based methods and machine learning, where a linear regressor is trained for each similarity type. The high results obtained by our system suggests that this technology is close to practical applications.

This article is structured as follows. The next section describes Europeana, the digital library used in this study. Section 3 introduces the types of similarity that we used in this work and the method to gather and annotate the pairs of items that comprise our dataset. Section 4 presents some discussion and analysis of the data set. Section 5 presents the tools used to generate similarity scores, followed by Section 6, which presents the systems that return typed similarity scores. Evaluation is described in Section 7, including a comparison to the state-of-the-art systems. Finally, Section 8 presents conclusions and future work.

2 Europeana

Europeana[5] is a web-portal that acts as a gateway to collections of cultural heritage items provided by a wide range of European institutions. It currently provides access to over 20 million digital records describing paintings, films, books, archival records and museum objects. The items are provided by around 1,500 institutions which range from major institutions, including the Rijksmuseum in Amsterdam, the British Library in London and the Louvre in Paris, to smaller and specialized organisations such as local museums. It therefore contains an aggregation of digital content from several sources and is not connected with any one physical museum.

Europeana stores the metadata about each item in an XLM-based format based on the Dublin Core standard. Information stored in this metadata includes a title (`<dc:title>`) and description (`<dc:description>`) for the artefact. There may also be information about the artefact's creator (e.g. painter, sculptor or photographer), stored in the `<dc:creator>` field, and date of creation, stored in the `<dc:date>` field. The date may be a specific date (e.g. 5th November 1905) or a time period (e.g. Bronze Age). The `<dc:collection>` field provides information about the collection the item came from (e.g. Kirklees Image Archive). Finally, cataloging information is provided for some items in the `<dc:subject>` field. This contains information about the item from a controlled vocabulary such as Library of Congress Subject Headings[6] or the Art and Architecture Thesaurus[7]. An example of metadata in the format used within Europeana is shown in Figure 1.

The metadata are created by different content providers and vary significantly across artefacts. Many of the items have only limited information associated with them, for example a very brief title. There is significant variation in the amount of information provided for some fields. For example, for some artefacts the `<dc:description>` field contains over a thousand words of text while for others it is empty. In addition, the content providers that contribute to Europeana use different controlled vocabularies and it is not straightforward to establish correspondences between them. Some providers do not make any use of controlled vocabularies so there is no information in the `<dc:subject>` field for many items. This variation in the information available makes the problem of determining the similarity between items quite challenging.

3 A dataset for typed similarity

This section describes the construction of a manually annotated data set for typed similarity generated from Europeana. The dataset is freely available[8].

[5] http://http://www.europeana.eu

[6] http://authorities.loc.gov/

[7] http://www.getty.edu/research/tools/vocabularies/aat/

[8] http://ixa2.si.ehu.es/sts

```
<dc:title>toy coins, crown (coin), toy coins</dc:title>
<dc:creator>The Fitzwilliam Museum, Cambridge, UK</dc:creator>
<dc:subject>Victoria (1837-1901) crown (coin) toy coins</dc:subject>
<dc:description>Artist:  Victoria (1837-1901), ruler - Queen of Great
Britain 1837-1901; Date(s):  1887 - 1901; Classification(s):  toy
coins, crown (coin), toy coins; Acquisition:  given by Withers, Paul,
2003-11-25 [CM.2666-2003]</dc:description>
```

Figure 1: Example of information about an artefact available in Europeana

3.1 Defining similarity types

The importance of typed-similarity was identified as part of PATHS[9], a research project on the development of exploratory search interfaces for cultural heritage collections, including Europeana (Agirre et al., 2013a). The interface developed by the project provided information about similar items in collections and recommendations about items a user might like to consult. Users of the system requested more information about why items were considered similar. Consequently we explored methods for generating information about the type of similarity that could be presented to the user. Discussions with users and analysis of the collection revealed seven types of similarity:

1. **similar author/creator** such as paintings by the same artist

2. **similar people involved** such as items showing the same people

3. **similar time period** such as items from the same year

4. **similar location** such as items showing the same place (e.g. a photograph and painting of the White House)

5. **similar event or action involved** such as items showing weddings, or people eating ice cream

6. **similar subject** such as items related to the same subject, e.g. horses

7. **similar description** items which have a similar descriptions

In addition, we also include a *general* similarity type which the annotators can choose when none of the seven types appears appropriate.

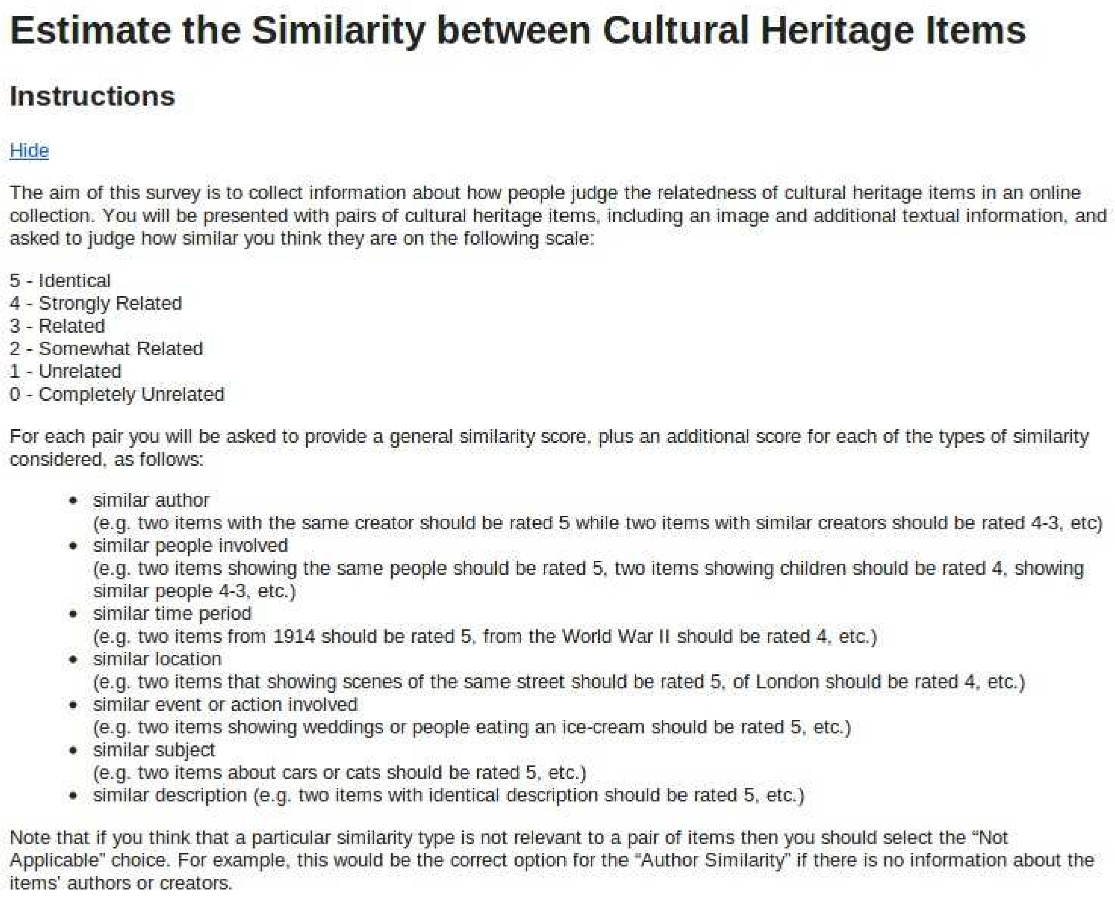
3.2 Selecting item pairs

Pairs of items were selected semi-automatically from Europeana. 25 pairs of items were manually selected for each of the seven similarity types (excluding general similarity), generating a total of 175 pairs. After removing duplicates and cleaning the dataset, 163 of these pairs remained. These manually selected pairs were then used as seeds to automatically select new pairs. The Europeana API was used to identify items the were similar to the seeds. For each seed, we created two chains of similar item pairs using an iterative process. The first chain of pairs was obtained using the current seed and a randomly chosen similar item from those provided by the Europeana API[10]. The newly identified item was then used as a new seed to continue building the chain of similar pairs. Thus, at each step, we obtained a new pair of similar items at *distance one*. The second chain followed the same iterative process, but selecting as new similar item among those appearing at *distance two* of the current seed in the chain. For each chain, we repeated the process up to five times.

This process yields 1500 pairs, the 163 that were manually selected, 892 from *distance one* chains and 445 from *distance two* chains. We then divided the data into training and testing sets containing 750 pairs each. The training data contains 82 manually selected pairs, 446 pairs from *distance one* chains and 222 pairs with from *distance two* chains. The

[9] `http://www.paths-project.eu`

[10] The Europeana API uses logs and textual descriptions to find similar items.



Estimate the Similarity between Cultural Heritage Items

Instructions

Hide

The aim of this survey is to collect information about how people judge the relatedness of cultural heritage items in an online collection. You will be presented with pairs of cultural heritage items, including an image and additional textual information, and asked to judge how similar you think they are on the following scale:

5 - Identical
4 - Strongly Related
3 - Related
2 - Somewhat Related
1 - Unrelated
0 - Completely Unrelated

For each pair you will be asked to provide a general similarity score, plus an additional score for each of the types of similarity considered, as follows:

- similar author
  (e.g. two items with the same creator should be rated 5 while two items with similar creators should be rated 4-3, etc)
- similar people involved
  (e.g. two items showing the same people should be rated 5, two items showing children should be rated 4, showing similar people 4-3, etc.)
- similar time period
  (e.g. two items from 1914 should be rated 5, from the World War II should be rated 4, etc.)
- similar location
  (e.g. two items that showing scenes of the same street should be rated 5, of London should be rated 4, etc.)
- similar event or action involved
  (e.g. two items showing weddings or people eating an ice-cream should be rated 5, etc.)
- similar subject
  (e.g. two items about cars or cats should be rated 5, etc.)
- similar description (e.g. two items with identical description should be rated 5, etc.)

Note that if you think that a particular similarity type is not relevant to a pair of items then you should select the "Not Applicable" choice. For example, this would be the correct option for the "Author Similarity" if there is no information about the items' authors or creators.

Figure 2: Annotation instructions.

test data follows a similar distribution.

Table 1 shows descriptive statistics for the six fields provided to the participants (number of non-empty fields, average length of field in tokens and standard deviation of field length). These statistics were computed from the 1500 items (750 pairs) in the training portion of the data set. A similar distribution was observed for the test set.

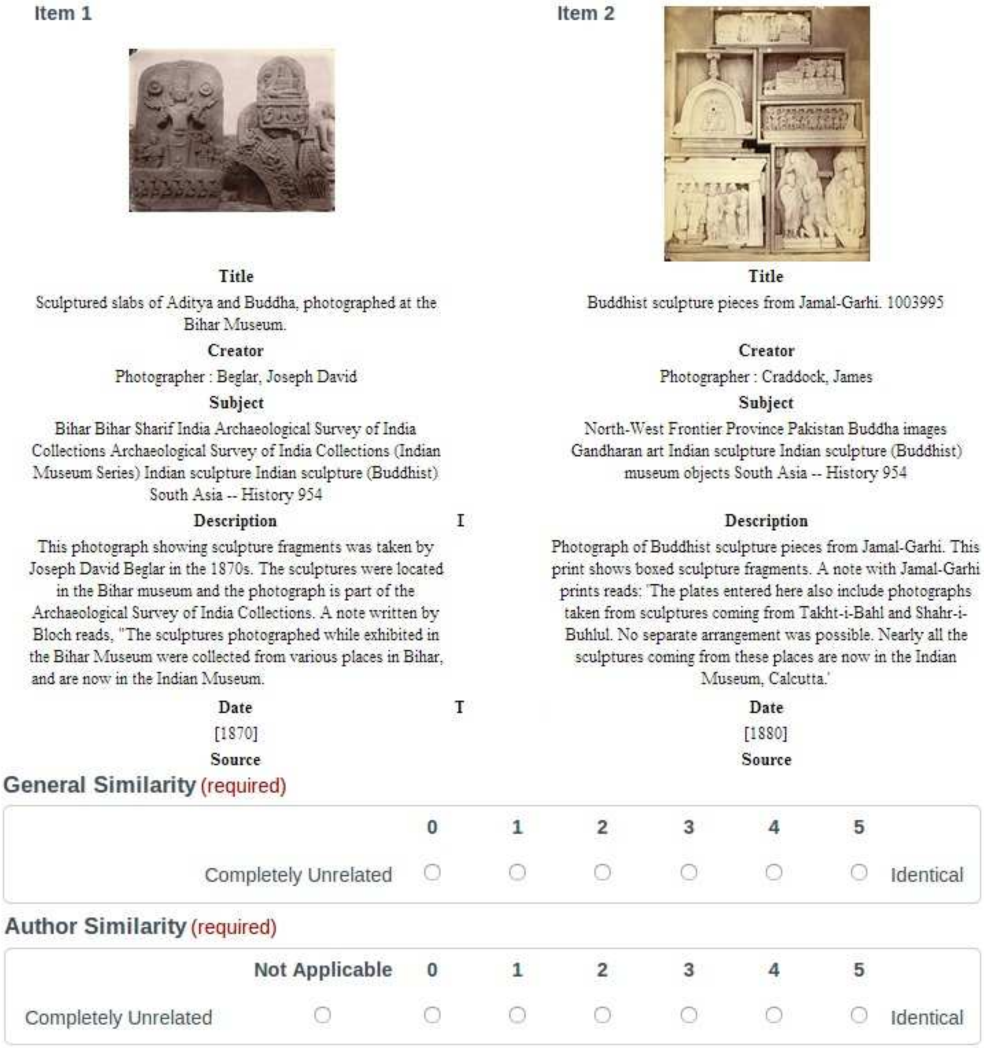
3.3 Annotation

The dataset was annotated using CrowdFlower[11], an online crowdsourcing platform. A survey was created containing the 1,500 pairs of the dataset (750 for training and 750 for testing). A set of 20 "gold" pairs with known answers were added for quality control[12]. Each annotator was initially shown four gold questions at the beginning for training, and then one gold question every two or four questions depending on the accuracy. If the accuracy for a particular annotator dropped to less than 66.7% percent, the survey was stopped and the answers for that annotator discarded. Each annotator was allowed to rate a maximum of 20 pairs to avoid annotators becoming tired or bored. To ensure quality, the task was restricted to annotators from a set of English speaking countries: UK, USA, Australia, Canada and New Zealand. Each pair of items included eight questions regarding different types of similarity (see below) and was annotated at least by 5 annotators. A total of $1,584$ annotators took part in the survey.

Figure 2 is a screenshot of the instructions provided to the annotators. Figure 3 shows how a a pair of items from the dataset is presented to the annotators. Annotators were asked to rate the similarity between pairs of cultural heritage items in the range 0 to 5. A *Not Applicable* option was also included to avoid annotators being forced to make a choice

[11] http://www.crowdflower.com/

[12] The gold pairs were chosen from those pairs manually selected by the authors.



Item 1

Title

Sculptured slabs of Aditya and Buddha, photographed at the Bihar Museum.

Creator

Photographer : Beglar, Joseph David

Subject

Bihar Bihar Sharif India Archaeological Survey of India Collections Archaeological Survey of India Collections (Indian Museum Series) Indian sculpture Indian sculpture (Buddhist) South Asia -- History 954

Description

This photograph showing sculpture fragments was taken by Joseph David Beglar in the 1870s. The sculptures were located in the Bihar museum and the photograph is part of the Archaeological Survey of India Collections. A note written by Bloch reads, "The sculptures photographed while exhibited in the Bihar Museum were collected from various places in Bihar, and are now in the Indian Museum.

Date

[1870]

Source

Item 2

Title

Buddhist sculpture pieces from Jamal-Garhi. 1003995

Creator

Photographer : Craddock, James

Subject

North-West Frontier Province Pakistan Buddha images Gandharan art Indian sculpture Indian sculpture (Buddhist) museum objects South Asia -- History 954

Description

Photograph of Buddhist sculpture pieces from Jamal-Garhi. This print shows boxed sculpture fragments. A note with Jamal-Garhi prints reads: 'The plates entered here also include photographs taken from sculptures coming from Takht-i-Bahl and Shahr-i-Buhlul. No separate arrangement was possible. Nearly all the sculptures coming from these places are now in the Indian Museum, Calcutta.'

Date

[1880]

Source

General Similarity (required)

	0	1	2	3	4	5	
Completely Unrelated	○	○	○	○	○	○	Identical

Author Similarity (required)

	Not Applicable	0	1	2	3	4	5	
Completely Unrelated	○	○	○	○	○	○	○	Identical

Figure 3: Pair of items as shown in the survey to annotators. Only *general* and *author* similarity types are displayed here. The annotators would see all types.

when they were unsure. In those cases the similarity score was calculated using the values provided by the other annotators (or 0 if there were no other annotators for a particular item).

3.4 Quality of annotation

To assess annotation quality, we compute the Pearson product-moment correlation of each annotator against the average of the rest of the annotators, as in (Grieser et al., 2011; Aletras et al., 2012). We then averaged all the correlations. This measure is identical to the one used for evaluation (see Section 7.1) and can be used to put those results into context. The inter-tagger correlation in the dataset for each type of similarity is as follows:

- General: 77.0
- Author: 73.1
- People Involved: 62.5
- Time period: 72.0
- Location: 74.3
- Event or Action: 63.9
- Subject: 74.5
- Description: 74.9

The correlation figures are high, with an average of 71.5, confirming that the task was well designed. The weakest correlations are for the *People Involved* and *Event or Action* types, suggesting they are the most difficult to identify. Other annotations exercises which use a similar method to gather similar-

Field	Non-empty	Avg. Length	Std. Dev.
Title	1500	5.9	4.5
Creator	1049	3.6	2.3
Subject	1434	7.8	7.4
Description	1469	77.0	169.4
Date	295	1.4	0.5
Source	21	1.3	0.9

Table 1: Corpus statistics for each of the fields in the training dataset.

ity annotations report comparable figures for inter-tagger agreement (Agirre et al., 2012).

We also computed confusion matrices for each of the similarity types (see Figure 4). The GENERAL, SUBJECT and DESCRIPTION similarity fields (Figures 4a, 4g and 4h) show most of the weight in the 0-0 and 5-5 cells, indicating that there is a lot of agreement between annotators when they judge pairs as 0 or as 5. Almost all the disagreement is on 4-5 and 5-4 cells (i.e. very close disagreement).

The pattern is slightly different for the other similarity types (Figures 4b, 4c, 4d, 4e and 4f). In addition to the weight in the 0-0 and 5-5 cells there is also a lot of weight in the 0-5 and 5-0 cells. To discover the reason for this we manually examined a subset of the 0-5 and 5-0 disagreements. We found that they were mainly caused by one of the annotators ignoring the information in the description. A typical case would be two items with the same author where one of the items did not have a `dc:creator` field, but which mentioned who the author was in the description. The annotator who ignored the text in the description would assign a pair 0, while the annotator who had read the description would assign it a 5. Other than that we can conclude that annotators agree most of the time. As in the previous case, the fine-grained disagreement is also concentrated on the 4-5 and 5-4 cells for these similarity types.

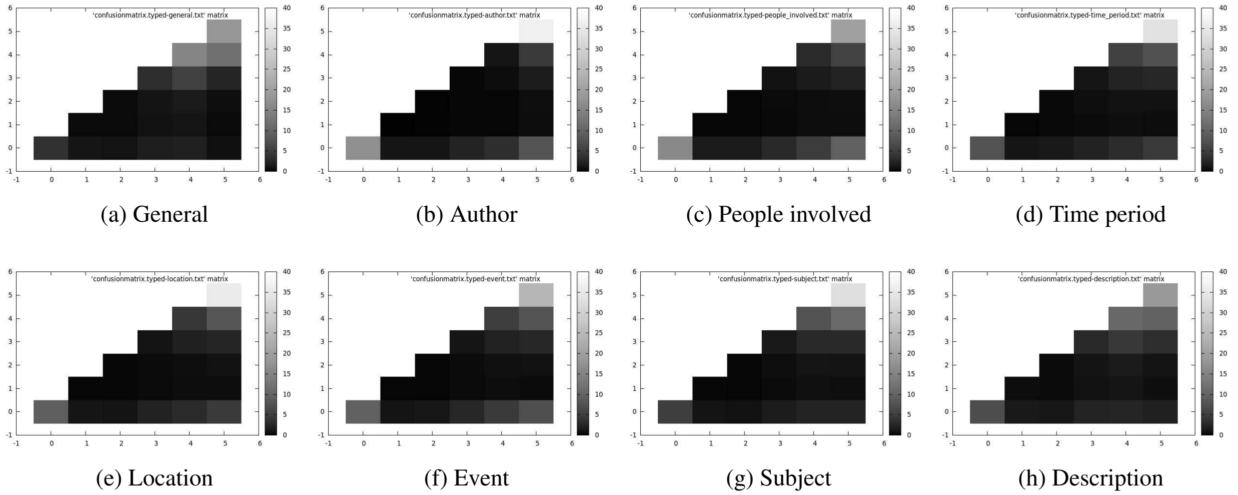
Figures 5, 6 and 7 show the average score value distribution, as assigned by the annotators, separated into five ranges. The majority of pairs are very closely related with nearly half of the pairs in the [4-5] range. (The EVENT and PEOPLE INVOLVED similarity types are exceptions which exhibit smoother distributions.) The dataset is skewed towards higher similarity scores since our aim was to select similar pairs of items rather than dissimilar ones.[13]

4 Discussion and analysis

We carried out an analysis of the annotations focussing on those pairs of items and annotation types where the annotators disagreed most. For instance, in the case of photographs (which form a substantial subset of the collection), there appears to be some confusion about the target of the annotation, specifically in relation to the AUTHOR similarity type. In these cases it is not clear whether the author type refers to the photographer who took the photograph or the creator of the item shown in the photograph (monument, building, painting, etc.). The same thing also happened for other types like PEOPLE INVOLVED in photographic items. Figure 8 shows an example of a pair of items where it is not clear if the annotation refers to the object in the picture or to the photograph itself.

Another source of disagreement is the poor quality of metadata. For instance, the CREATOR field might contain the institution that keeps the item (e.g. *Fitzwilliam Museum*), a generic term (e.g. *staff*) or even just *none*. Some annotators assign the maximum score to cases where the term is the same for both items, while others read the description of the items, which specifies the author or indicates that it is unknown, and score the pair accordingly. Figure 9 shows a pair of items where the metadata indicates *staff* as creators and the description contains the actual author (designer) of both items (*Sir Bernard de Gomme*). One of the annotators, seeing that the metadata was not useful, rated the author similarity as *Not Applicable* (NA), while the rest did read the

[13] Pearson is known to have issues when distributions are skewed. We checked the inter-tagger correlations using a down-sampled version of the full data, and the inter-tagger correlations we obtained were slightly higher.



(a) General

(b) Author

(c) People involved

(d) Time period

(e) Location

(f) Event

(g) Subject

(h) Description

Figure 4: Confusion matrices for the eight similarity types.

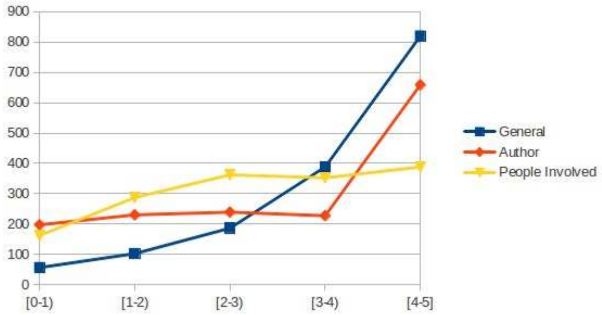


Figure 5: Score value distribution, as assigned by annotators, for general, author and people fields.

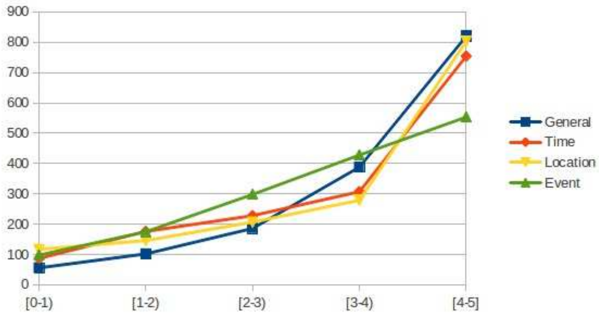


Figure 6: Score value distribution, as assigned by annotators, for general, time period, location and event fields.

description and rated the author similarity accordingly. Taking the average produced 4 as the final value in the gold standard.

In another example (Figure 10) we can see the metadata for a pair of items, photograph of (different) bridges. The creator field lists *unknown* in one case and the author of the photograph in another (*Eric de Mare* is a well-known British photographer), but the description explicitly mentions the builders of each bridge: *John Rennie* for one, and his two sons, *George* and *John*, for the other. The scores provided by the annotators of the *author* similarity is 2, 3, 3, 0 and 0. In this case, it seems that the last two annotators have not read the full description of the items, while the first three did recognise that the authors of both bridges are related but not the same.

In order to explore the effect of incorrect or incomplete metadata on the annotation process we studied annotators' behaviour while completing the task. We enrolled some Ph.D. students and asked them to annotate some of the conflicting pairs. We directly observed the annotators as they completed the task and also interviewed them after they had completed it. The study showed that the order of the fields and questions effected the annotations. For instance, the annotators rated the *author* similarity before the *description* similarity. In the absence of metadata in the *author* field some of the students evaluated this similarity as 0, without checking the *description*. They later identified the *author* in the *description* field, but some tended not to alter the score that has already been assigned for *author* similarity. This study suggests that annotators can be

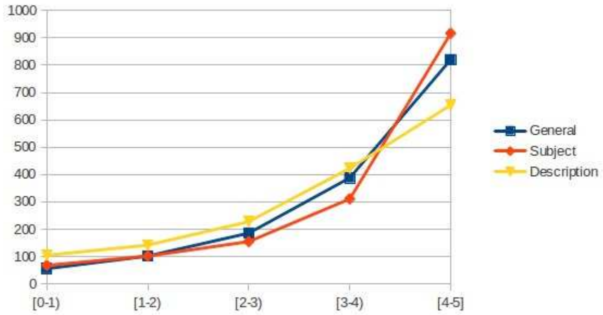


Figure 7: Score value distribution, as assigned by annotators, for general, subject and description fields.

confused by incorrect or incomplete metadata. For any future annotation exercises it would make sense to control the order in which the metadata is presented to the annotators so that the description field is presented early (just after the title) since it provides the most general description for the item in most cases.

Overall our analysis suggests that although the quality of the annotation is very good, it may also be possible to improve it further. For instance, clarifying the photograph vs. item issue for the AUTHOR type and by providing specific instructions in face of poor quality metadata in order to pay more attention to the text in the description.

5 Similarity methods

In this section we present the methods used for computing similarity and to build the typed-similarity systems described in the next section. Those tools are Bag-of-Words similarity using TF.IDF (Section 5.1), LDA (Section 5.2), the Wikipedia Link Vector Model (Section 5.3) and random walks over WordNet and Wikipedia graphs (Section 5.4). Each of these methods provide a different technique that can be applied to compute the similarity between a pair of texts.

5.1 TF.IDF

A common approach for computing similarity between texts is to represent the documents as a Bag-Of-Words (BOW). Each BOW is a vector consisting of the words contained in the document in which each dimension corresponds to a word and the weight is the frequency with which the word occurs within the document. The similarity between two documents can be computed as the cosine of the angle between their vectors. If two documents are identical the cosine value of their vectors is 1 while if they share no common terms the cosine value is 0.

This approach is usually improved by giving more weight to words which occur in few documents and less weight to common words which tend to occur in many documents (e.g. *the*). We used the Inverse Document Frequency (IDF) (Baeza-Yates and Ribeiro-Neto, 1999) using counts from the Culture Grid collection[14] in order to weight words. Thus, the **TF.IDF** similarity between items a and b is defined as follows:

$$sim_{\text{tf.idf}}(a,b) = \frac{\sum_{w\in a,b} \text{tf}_{w,a} \times \text{tf}_{w,b} \times \text{idf}_w^2}{\sqrt{\sum_{w\in a}(\text{tf}_{w,a} \times \text{idf}_w)^2} \times \sqrt{\sum_{w\in b}(\text{tf}_{w,b} \times \text{idf}_w)^2}}$$

where $\text{tf}_{w,x}$ is the frequency of the term w in $x \in \{a,b\}$ and idf_w is the inverted document frequency of the word w.

5.2 LDA

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a statistical method that learns a set of latent variables, called topics, describing the contents of a document collection. Given a topic model, documents can be viewed as a set of probability distributions over topics, θ. The distribution for an individual document i is denoted as θ_i.

The similarity between a pair of texts is estimated by comparing their topic distributions (Aletras et al., 2012; Aletras and Stevenson, 2012). This is achieved by considering each distribution as a vector (consisting of the topics corresponding to an item and its probability) then computing the cosine of the angle between them, i.e.

$$sim_{LDA}(a,b) = \frac{\vec{\theta_a} \cdot \vec{\theta_b}}{|\vec{\theta_a}| \times |\vec{\theta_b}|}$$

where $\vec{\theta_x}$ is the vector created from the probability distribution generated by LDA for text x.

To implement this approach an LDA model consisting of 100 topics was trained using the *gensim*

[14] Culture Grid (`http://www.culturegrid.org.uk/`) is the digital content provider service from the Collection Trust and forms part of Europeana. It contains information about over one million items.

	Similarity type	
Title	**General:**	**Title**
Sculptured slabs of Aditya and Buddha, photographed at the Bihar Museum.	2.2	Buddhist sculpture pieces from Jamal-Garhi. 1003995
Creator	**Author:**	**Creator**
Photographer : Beglar, Joseph David	1.6	Photographer : Craddock, James
Subject	**Subject:**	**Subject**
Bihar Bihar Sharif India Archaeological Survey of India Collections Archaeological Survey of India Collections (Indian Museum Series) Indian sculpture Indian sculpture (Buddhist) South Asia -- History 954	2.6	North-West Frontier Province Pakistan Buddha images Gandharan art Indian sculpture Indian sculpture (Buddhist) museum objects South Asia -- History 954
Description	**Description:**	**Description**
This photograph showing sculpture fragments was taken by Joseph David Beglar in the 1870s. The sculptures were located in the Bihar museum and the photograph is part of the Archaeological Survey of India Collections. A note written by Bloch reads, "The sculptures photographed while exhibited in the Bihar Museum were collected from various places in Bihar, and are now in the Indian Museum. A paper by Mr Broadley, dealing with this collection, was published in Journal of the Asiatic Society of Bengal, vol. XLI, part I, 1872, pp. 209-311." Aditya is depicted on the left slab whilst Buddha can be seen in a reclining position on the right sculpture. There are also two architectural sculptures shown in the photograph .	2	Photograph of Buddhist sculpture pieces from Jamal-Garhi. This print shows boxed sculpture fragments. A note with Jamal-Garhi prints reads: 'The plates entered here also include photographs taken from sculptures coming from Takht-i-Bahl and Shahr-i-Buhlul. No separate arrangement was possible. Nearly all the sculptures coming from these places are now in the Indian Museum, Calcutta.'
Date	**Time period:**	**Date**
[1870]	2.8	[1880]

Figure 8: Sample pair of items, where it is not clear whether the annotators need to refer to the items in the photographs, or to the photographs themselves. For each item, the contents of the fields in the metadata are shown. In the center of the figure, the gold standard scores for each of the types is given.

	Similarity type	
Title	**General:**	**Title**
Tilbury Fort, Tilbury, Essex	3.6	The Royal Citadel, Plymouth, Devon
Creator	**Author:**	**Creator**
staff	4	staff
Subject	**Subject:**	**Subject**
Aerial View Fort	3.8	Aerial View Coastal Fort Military
Description	**Description:**	**Description**
Tilbury Fort was designed in 1670 by Charles II's chief engineer, Sir Bernard de Gomme, in response to Dutch raids on the Thames. It is one of the best surviving examples of continental-inspired bastion defence in Britain.	3.4	Built between 1665-1667 on the site of Plymouth Fort, the Royal Citadel was designed by Dutch military engineer Bernard de Gommes to protect Plymouth from an attack by his own countrymen. More than three centuries later, the Citadel continues its military traditions and is now home to the British Army's 29 Commando Regiment.
Date	**Time period:**	**Date**
[1993]	4.2	[1999]
Source	**Location:**	**Source**
	3.4	

Figure 9: Sample of a pair of items which contain poor metadata in the author field (images removed for space). For each item, the contents of the fields in the metadata are shown. In the center of the figure, the gold standard scores for each of the types is given.

	Similarity type	
Title	**General:**	**Title**
London Bridge, City of London	3.2	Serpentine Bridge, Hyde Park, Westminster, Greater London
Creator	**Author:**	**Creator**
not known	1.6	de Mare, Eric
Subject	**Subject:**	**Subject**
	3	Waterscape Animals Bridge Gardens And Parks
Description	**Description:**	**Description**
A view of London Bridge which is packed with horse-drawn traffic and pedestrians. This bridge replaced the earlier medieval bridge upstream. It was built by John Rennie in 1823-31. A new bridge, built in the late 1960s now stands on this site today.	2.2	The Serpentine Bridge in Hyde Park seen from the bank. It was built by George and John Rennie, the sons of the geat architect John Rennie, in 1825-8.
Date	**Time period:**	**Date**
	4	[1945, 1980]

Figure 10: Sample of a pair of items which contain contradictory authorship information (images removed for space). For each item, the contents of the fields in the metadata are shown. In the center of the figure, the gold standard scores for each of the types is given.

package[15] with hyperparameters (α, β) were set to $1/num\_of\_topics$.

5.3 WLVM

An algorithm described by Milne and Witten (2008) associates Wikipedia articles with a document using machine learning techniques. We make use of that method to represent each item as a set of Wikipedia articles. The similarity of two documents can be thus computed as a function of the similarity between the Wikipedia articles associated with each text. We measured the similarity between Wikipedia articles using the Wikipedia Link Vector Model (WLVM) (Milne, 2007), which uses both the link structure and the article titles. Each link is weighted by the probability of its occurrence. Thus, the value of the weight w for a link $x \to y$ between articles x and y is:

$$w(x \to y) = |x \to y| \times \log \left(\sum_{z=1}^{t} \frac{t}{z \to y} \right)$$

where t is the total number of articles in Wikipedia. The similarity of articles is compared by forming vectors of the articles which are linked from them and computing the cosine of their angle. For example the vectors of two articles x and y are:

$$x = (w(x \to l_1), w(x \to l_2), ..., w(x \to l_n))$$
$$y = (w(y \to l_1), w(y \to l_2), ..., w(y \to l_n))$$

where x and y are two Wikipedia articles and $x \to l_i$ is a link from article x to article l_i.

The similarity between two documents can then be computed by performing pairwise comparison between the corresponding articles using WLVM, selecting the highest similarity score for each, as follows:

$$sim(a,b) = \frac{1}{2} \left(\frac{\sum_{w_1 \in a} \arg\max_{w_2 \in b} WLVM(w_1, w_2)}{|a|} + \frac{\sum_{w_2 \in b} \arg\max_{w_1 \in a} WLVM(w_2, w_1)}{|b|} \right)$$

where a and b are two texts, $|a|$ the number of Wikipedia articles in a and $WLVM(w_1, w_2)$ is the WLVM similarity between articles w_1 and w_2.

5.4 Random walks

Random walks have been successfully used to compute the similarity between words (Agirre et al., 2010) and we extended these techniques to compute similarity between documents. We used the semantic disambiguation and similarity algorithm UKB[16] (Agirre and Soroa, 2009), which applies personalized PageRank on a graph generated from the English WordNet (Fellbaum, 1998), or alternatively, from Wikipedia.

To compute similarity between two words using UKB, we first represent WordNet as a graph $G = (V, E)$: graph nodes represent WordNet concepts (synsets) and dictionary words; relations among synsets are represented by undirected edges; and dictionary words are linked to the synsets associated to them by directed edges. We used the graph provided by UBK package. We then compute the personalized PageRank over WordNet separately for each of the words, producing two vectors with the probability distribution over WordNet synsets. The similarity between the words can be computed as the cosine between the two probability distributions.

The similarity between two documents can be computed initializing the random walks using the words in the respective texts to obtain a vector of probability distribution over synsets, and computing the cosine.

In addition to WordNet, we also used the Wikipedia graph, where the nodes correspond to Wikipedia articles, and the edges to hyperlinks between articles. We used version 3.0 of WordNet and the publicly available dump of Wikipedia dated 25th of May of 2011.

6 System construction

In this section we introduce our systems for identifying typed similarity. We first explain how the text in the items was processed, followed by descriptions of the three systems we implemented, a baseline approach, knowledge-based approach and machine learning system.

6.1 Processing text in the items

The text in the items was pre-processed using Stanford CoreNLP (Finkel et al., 2005; Toutanova et al.,

[15] http://pypi.python.org/pypi/gensim

[16] http://ixa2.si.ehu.es/ukb/

```
<entity netype="ORG" lemma="Fitzwilliam_Museum" field="dc:creator"/>
<entity netype="LOC" lemma="Cambridge" field="dc:creator"/>
<entity netype="LOC" lemma="UK" field="dc:creator"/>
<entity netype="LOC" lemma="Victoria" field="dc:subject"/>
<entity netype="DATE" lemma="1837-1901" field="dc:subject"/>
<entity netype="LOC" lemma="Victoria" field="dc:description"/>
<entity netype="DATE" lemma="1837-1901" field="dc:description"/>
<entity netype="LOC" lemma="Great_Britain" field="dc:description"/>
<entity netype="DATE" lemma="1837-1901" field="dc:description"/>
<entity netype="DATE" lemma="1887_-_1901" field="dc:description"/>
<entity netype="PER" lemma="Withers" field="dc:description"/>
<entity netype="PER" lemma="Paul" field="dc:description"/>
<entity netype="DATE" lemma="2003-11-25" field="dc:description"/>
```

Figure 11: Example of NER analysis on the item shown in Figure 1

2003), including tokenization, part-of-speech tagging, named entity recognition and classification (NERC) and date detection. The NERC module is key, as it detects people and locations.

6.2 Baseline system

We implemented a baseline system using only TF.IDF-based similarity (see Section 5.1) to provide an indication of the performance that could be obtained using a simple approach. TF.IDF was applied differently for each similarity type.

- General: cosine similarity of **TF-IDF** vectors created using tokens from all fields.
- Author: cosine similarity of **TF-IDF** vectors created using dc:Creator field.
- People involved, time period and location: cosine similarity of **TF-IDF** vectors created from people, locations and date expressions recognized by NERC in all fields. Figure 11 shows a sample of the people, locations and dates which were automatically detected in the metadata for the item in Figure 1.
- Events: cosine similarity of **TF-IDF** vectors constructed from verbs in all fields.
- Subject and description: cosine similarity of **TF-IDF** vectors created from respective fields.

6.3 Knowledge based approach

The second approach built on the baseline to make use of information from Wikipedia and WordNet (Section 5.4). Rather than applying TF.IDF similarity to all fields, as the baseline system did, different processes were applied to each field:

- Author: similarity using random walks on Wikipedia for the person entities in the dc:Creator field.
- People involved: similarity using random walks on Wikipedia for the person entities recognized by NERC in all fields.
- Location: similarity using random walks on Wikipedia for the location entities recognized by NERC in all fields.
- Events: similarity using random walks on WordNet for event verbs and nouns in all fields. A list of verbs and nouns that may denote events was derived using morphosemantic links in WordNet[17].

Results on the training data showed that the coverage of random walks for the aforementioned fields was quite low (except for event similarity, where good performance was obtained). This was caused by the large number of cases where the Stanford parser did not find entities which were in Wikipedia. Consequently the scored returned by the random walks were combined with the TF.IDF similarity scores presented in Section 6.2 as follows: if UKB similarity returns a score then it is multiplied with the TF.IDF score, otherwise we return the square of

[17] `http://wordnetcode.princeton.edu/standoff-files/morphosemantic-links.xls`

the TF.IDF similarity score.

In addition, the general similarity was improved in two ways: lemmas were used instead of word forms and Wikipedia was used to compute IDF scores instead of the CultureGrid collection. (We found that using CultureGrid lead to some undesirable outcomes, e.g. the word *coin* had a very low IDF because it occurs very frequently in the CultureGrid collection.)

Finally, a dedicated similarity measure for dates was devised, in order to model that, e.g. 1500 and 1550 are similar dates while 99 and 1999 are not. To measure the time similarity between a pair of items we first need to identify the time expressions contained in both items. We assume that the year of creation or the year denoting when the event referenced by an item took place are good indicators of temporal similarity. Information about years mentioned in each item's meta-data is extracted using the following pattern: $[1|2][0-9]\{3\}$. Using this approach, each item is represented as a set of numbers denoting the years extracted from the item.

Time similarity between two items is computed based on the similarity between their associated years. Similarity between two years is defined as:

$$sim_{year}(y_1, y_2) = max\{0, 1 - |y1 - y2| * k\}$$

where k is a parameter to weight the difference between two years, e.g. for $k = 0.1$ all items that have difference of 10 years or more are assigned a score of 0. We experimented with various values for K and obtained the best results for $k = 0.1$.

Finally, time similarity between items a and b is computed as the maximum of the pairwise similarity between their associated years:

$$sim_{time}(a, b) = max_{\substack{\forall i \in a \\ \forall j \in b}}\{0, sim_{year}(a_i, b_j)\}$$

The, we substituted the preliminary TIME similarity score of the baseline system by the measure obtained using the method presented in this section.

6.4 Machine learning system

The systems described so far used dedicated similarity measures to model each similarity type separately. In some cases, we are able to provide more than one option for each type of similarity. The machine learning system takes each of those similarity measures as features and uses linear regression (from Weka (Hall et al., 2009)) to learn models that fit those features to the training data.

We used further similarity scores as features for general similarity, including LDA (Section 5.2) and WLVM (Section 5.3). In addition, we used random walks (Section 5.4) to generate a probability distribution over WordNet synsets for all of the words in each item. Similarity between two words is computed by creating vectors from these distributions and comparing them using the cosine of the angle between the two vectors. If a words does not appear in WordNet its similarity value to every other word is set to 0.

The similarity between a pair of items is computed by performing pairwise comparison between the words they contain and selecting the highest similarity score. The approach is similar to the one used to identify the similarity between a pair of texts based on their WLVM scored described in Section 5.3.

7 Evaluation

This Section describes evaluation of the typed similarity systems described previously. It presents the evaluation metrics used, results obtained during the development phase (using the training portion of the dataset) and the final results obtained using the test data. Results are compared against state of the art systems. Note that we follow the same partition of training and test data that was used in the *SEM 2013 shared task (see Section 7.4) making the results directly comparable.

7.1 Evaluation metric

System performance is evaluated by computing the Pearson product-moment correlation between the scores returned by the systems and the gold standard values (Rubenstein and Goodenough, 1965), an approach often employed in word similarity experiments. Statistical significance between results is computed using a one-tailed parametric test based on Fisher's z-transformation (Press et al., 2002, equation 14.5.10).

7.2 Development

The training data was used to develop the systems and check performance. Results for the machine learning system were generated using 10-fold cross-validation.

Type	Feature	Results	Δ Baseline
General	Baseline	65.8	-
	LDA	68.0	2.2
	TF-IDF$_{Wiki}$	**72.7**	**6.9**
	UKB$_{Wiki}$	54.1	-11.7
	WLVM	56.1	-9.7
Author	Baseline	39.6	-
	UKB$_{Wiki}$	27.2	-12.4
	Combined UKB$_{Wiki}$	44.7	**5.1**
People involved	Baseline	**47.4**	-
	UKB$_{Wiki}$	29.7	-17.7
	Combined UKB$_{Wiki}$	46.5	-0.9
Location	Baseline	47.2	-
	UKB$_{Wiki}$	22.2	-25.0
	Combined UKB$_{Wiki}$	**48.0**	**0.8**
Time	Baseline	54.8	-
	Improved Time Measure	**58.8**	**4.0**
Event	Baseline	26.4	-
	UKB$_{WN}$	**28.5**	**2.1**
	Combined UKB$_{WN}$	28.3	1.8
Subject	Baseline	49.8	-
Description	Baseline	53.9	-

Table 2: Development results on each similarity type for the Baseline approach (TF.IDF) and the improved components applied in the knowledge based approach (cf. Sections 6.2 and 6.3).

System	General	Author	People	Time	Location	Event	Subject	Description	Mean
Baseline	65.8	39.6	47.4	54.8	47.2	26.4	49.8	53.9	48.1
Knowledge	72.7	44.7	46.5	58.8	48.0	28.5	49.7	53.9	50.3
ML system	**78.7**	**69.4**	**69.7**	**76.5**	**74.9**	**65.5**	**75.9**	**80.7**	**73.9**

Table 3: Development results of each system for each type of similarity, including the mean of all types.

Table 2 shows the results obtained using the baseline system and improved components from the knowledge based system for each of the similarity types, including the improvement over the baseline. The results show that the use of Wikipedia counts when computing TF.IDF improve the results of general similarity, and yield the best results overall, with 6 absolute points of improvement over the baseline.

The use of random walks over Wikipedia (UKB$_{Wiki}$) leads to results that are worse than the baseline approach, unless both scores are combined. (The combined score was obtained by multiplying the individual scores. If one of the algorithms did not yield a score, we squared the score of the other algorithm.) When a combination is used results improve for *Author* and *Location*, but not for *People involved*. The use of random walks over WordNet (UKB$_{WN}$) for events does improve over the baseline, without need of combination.

The dedicated time similarity measure also improves the results over the baseline. Note that we did not experiment with any improvements for the subject and description fields given the strong results generated by the baseline system.

The results of the full systems on each individual type in the training data are shown on Table 3, together with the mean score across all types. The table shows that the Baseline system (**Baseline**) obtains the lowest results, with the knowledge based system (**Knowledge**) getting better results overall

Team and run	General	Author	People	Time	Location	Event	Subject	Description	Mean	Rank
Baseline	66.9	42.8	44.6	50.0	48.4	30.6	50.2	58.1	48.9	
Knowledge	72.6	45.7	44.7	57.6	48.6	30.9	50.2	58.1	51.0	6
ML system	**74.6**	**66.6**	**65.4**	**74.1**	**72.6**	**65.5**	**74.2**	**77.6**	**71.3**	**3**

Table 4: Test results of each system for each type of similarity, including the mean of all types.

and for most types (except for *People Involved*). Linear regression (**ML system**) improves results considerably for all types, yielding a mean value of 73.9. Values over 65 are obtained for all types, a values that is usually taken to mean a strong association.

7.3 Test results

Table 4 shows the results of our systems in the test dataset. The results are very similar to those obtained on the training data, but in this case the **Knowledge based system** performs better or equal to the baseline system in all types. The **Machine Learning system** provides the best results by far for all types, with correlations over 65 in all cases. The difference between the knowledge based system and baseline is not statistically significant, but the difference between the Machine Learning and knowledge based systems is (p-value $<$ 0.02).

The high correlations obtained by our machine leaning system suggest that deploying automatic systems for typed-similarity in real tasks is feasible. In fact, the correlations attained by our best system (see Table 4) are comparable to the inter-tagger correlations obtained during annotation (see Section 3.4).

7.4 Performance in Shared Task

The systems described in this article participated in the *SEM 2013 shared task (Agirre et al., 2013b). Our baseline system was used as the overall task baseline against which all runs were compared. This baseline system actually outperformed many of the submitted systems for various similarity types and achieved an overall ranking of 8th out of the 14 submitted systems. The knowledge based system was ranked in 6th place overall and the machine learning system in 3rd place. The best system (Croce et al., 2013) applied an approach that combined Support Vector Regression with compositional distributional semantics to achieve an overall mean score of 76.2 across all similarity types.

8 Conclusion and Future Work

This article introduced the new problem of typed similarity, determining the type of the relation the holds between pairs of similar items. Typed similarity has various applications including providing recommendations and improving search through collections.

The problem was investigated within a subset of a large digital library of cultural heritage items. Seven types of similarity specific to this domain were identified: author, time, location, involved people, events, subject and description. A data set was created using 1500 pairs of items and annotated using crowdsourcing. Analysis of the annotation revealed an average Pearson correlation of 71.5, this high inter-annotator agreement indicates that the task is well-defined.

Three approaches to automatically determining similarity type were explored. The simplest approach was used as a baseline against which knowledge based and machine learning approaches were compared. The best results were obtained using the machine learning system which employed linear regression. This approach yields a mean Pearson correlation of 71.3, close to the human performance for this task.

The task has been used as a community evaluation exercise, the *SEM 2013 shared task on Semantic Textual Similarity (Agirre et al., 2013b). The exercise attracted 14 system runs from 6 teams.

The typed similarity system presented here has been deployed within an exploratory search interface for Europeana (Agirre et al., 2013a). When users view an individual Europeana item in this system they are also shown up to 25 similar items together with the similarity type to provide a motivation for displaying particular items. The type of the similarity is determined automatically using the machine learning system.

In future, we would like to carry out further evaluation of this application to determine how useful users find this information within the application. In addition, we would like to explore the typed similarity problem in other domains, where a different set of similarity types are likely to be relevant.

Acknowledgements

The development of the typed-similarity dataset was supported by the PATHS project (http://paths-project.eu) funded by the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement no. 270082. This work is partially funded by the Spanish government, PhD grant FPU12/06243 of Aitor Gonzalez-Agirre, Skater project TIN2012-38584-C06-02. We thank Europeana and all contributors to Europeana for sharing their content through the API.

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece.

Eneko Agirre, Montse Cuadros, German Rigau, and Aitor Soroa. 2010. Exploring knowledge bases for similarity. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC10). European Language Resources Association (ELRA). ISBN: 2-9517408-6-7. Pages 373–377."*.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Eneko Agirre, Nikolaos Aletras, Paul Clough, Samuel Fernando, Paula Goodale, Mark Hall, Aitor Soroa, and Mark Stevenson. 2013a. Paths: A system for accessing cultural heritage collections. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 151–156, Sofia, Bulgaria, August. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013b. *sem 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Nikolaos Aletras and Mark Stevenson. 2012. Computing similarity between cultural heritage items using multimodal features. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 85–93, Avignon, France.

Nikolaos Aletras, Mark Stevenson, and Paul Clough. 2012. Computing similarity between items in a digital library of cultural heritage. *J. Comput. Cult. Herit.*, 5(4):16:1–16:19, December.

R. Baeza-Yates and B. Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison Wesley Longman Limited, Essex.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, March.

F. Bohnert, D. Schmidt, and I. Zuckerman. 2009. Spatial Process for Recommender Systems. In *Proc. of IJCAI 2009*, pages 2022–2027, Pasadena, CA.

J. Bowen and S. Filippini-Fantoni. 2004. Personalization and the Web from a Museum Perspective. In *Proc. of Museums and the Web 2004*, pages 63–78.

Danilo Croce, Valerio Storch, and Roberto Basili. 2013. Unitor-core_typed: Combining text similarity and semantic filters through sv regression. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 59–65, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. 2002. Placing Search in Context: The Concept Revisited. *ACM Trans. on Information Systems*, 20(1):116–131.

K. Grieser, T. Baldwin, and S. Bird. 2007. Dynamic Path Prediction and Recommendation in a Museum Environment. In *Proc. of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, pages 49–56, Prague, Czech Republic.

K. Grieser, T. Baldwin, F. Bohnert, and L. Sonenberg. 2011. Using Ontological and Document Similarity to Estimate Museum Exhibit Relatedness. *Journal of*

Computing and Cultural Heritage (JOCCH), 3(3):1–20.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

M. Hearst. 2009. *Search User Interfaces*. Cambridge University Press.

D. Jurafsky and J. Martin. 2009. *Speech and Language Processing*. Pearson, second edition.

C. Manning and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

G. Marchionini. 2006. Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4):41–49.

M. McCandless, E. Hatcher, and O. Gospodnetic. 2010. *Lucene in Action*. Manning Publications.

D. Milne and I. Witten. 2008. Learning to Link with Wikipedia. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM'2008)*, Napa Valley, California.

D. Milne. 2007. Computing semantic relatedness using Wikipedia's link structure. In *Proceedings of the New Zealand Computer Science Research Student Conference*.

M. O'Donnell, C. Mellish, J. Oberlander, and A. Knott. 2001. ILEX: An architecture for a dynamic hypertext generation system. *Natural Language Engineering*, 7:225–250.

I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. 2006. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*.

W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. 2002. *Numerical Recipes: The Art of Scientific Computing V 2.10 With Linux Or Single-Screen License*. Cambridge University Press.

P. Resnick and H. Varian. 1997. Recommender systems. *Communications of the ACM*, 40(3):56–58.

I. Roes, N. Stash, Y. Wang, and L. Aroyo. 2009. A personalized walk through the museum: the CHIP interactive tour guide. In *Proc. of the 27th International Conference on Human Factors in Computing Systems*, pages 3317–3322, Boston, MA.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633, October.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.